DOCUMENT RESUME

ED 420 676 TM 028 351

AUTHOR Kalohn, John C.; Spray, Judith A.
TITLE Effect of Item Selection on Item Exposure Rates within a Computerized Classification Test.
PUB DATE 1998-04-14
NOTE 19p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Diego, CA, April 13-16, 1998).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Certification; *Classification; *Computer Assisted Testing; Item Banks; *Licensing Examinations (Professions); *Selection; Test Construction; *Test Items
IDENTIFIERS *Item Exposure (Tests)

ABSTRACT

The purpose of many certification or licensure tests is to identify candidates who possess some level of minimum competence to practice their profession. In general, this type of test is referred to as classification testing. When this type of test is administered with a computer, the test is a computerized classification test (CCT). This paper addresses the effect of item selection on item exposure rates within a CCT. A testing program's ideal CCT item pool would consist of items that tended to measure the best at the latent equivalent of the passing score (theta p). For this study, it was hypothesized that there would be significant differences in item exposure rates for the extreme items in the pool (those that measured best at high and low values of theta), but that the overall impact of these differences would be negligible. A simulation was designed, using an actual item pool of 1,235 items. Observed item exposure rates were then calculated under two target exposure rate (TER) conditions, one set at 0.20 and the other at 0.10. Both methods of item selection produced about the same degree of classification accuracy, but performance of the CCTs in terms of classification accuracy, declined slightly for TER=0.10, as would be expected when the use of more items with less information is necessary. (Contains one table, seven figures, and six references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 420 676

Effect of Item Selection on Item Exposure Rates Within a Computerized Classification Test

John C. Kalohn

Judith A. Spray

ACT, Inc.

TM028351

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

John Kalohn

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Paper presented at the Annual Meeting of the National Council on Measurement in Education, April 14, 1998, San Diego, CA.

Effect of Item Selection on Item Exposure Rates Within a Computerized Classification Test

The purpose of many certification or licensure tests is to identify candidates who possess some level of minimum competency to safely practice their profession. These tests are similar to educational criterion-referenced or *mastery* tests in which examinees either pass the test and are classified as *masters*, or fail and are classified as *nonmasters*. In general we refer to this type of testing as *classification* testing; when the test is administered via computer so that test items are selected and administered according to some algorithm, the test is called a *computerized classification test* or CCT.

There are several approaches that can be used to implement a CCT. Many of these either suggest or require the test items making up the CCT *item pool* be calibrated and scaled on an IRT metric. Once calibrated and scaled, the items can be selected for possible administration either at an estimate of the candidate's current ability using responses to previously administered test items, or at the point on the latent scale which corresponds to the examination's passing score or decision threshold. For some CCTs, there may be more than one decision point. However, for the purposes of this paper, we will only address a single passing criterion.

Items are selected for possible administration based on one or more item characteristics, which usually can be distinguished or classified as either *psychometric* or *content-based.* A popular psychometric criterion, for example, is Fisher's information, and the selection criterion is the point on the latent scale where the item's information is

maximized (θ_{max}). Content-based characteristics usually refer to an item's test blueprint category or domain, where the test blueprint normally dictates what percentage of the items must come from each of the blueprint's domains.

Previous research has shown that if the primary purpose of the CCT is to make a single classification decision, the item's psychometric criterion should be the maximum information at the passing score. This decision ensures that the test will provide the most power and yield the least classification error (Reckase & Spray, 1994). The major implication of this finding is that a testing program's ideal CCT item pool would consist of items that tended to measure the best (i.e., had their maximum information values, θ_{max}) at the latent equivalent of the passing score, θ_p.

Unfortunately, most item pools from certification programs do not conform to this ideal. It has been our experience that many voluntary certification pools contain items that, in general, follow the shape of the distribution of the examinee population. In other words, a graph of each item's maximum information value plotted as a function of θ reveals more items clustering at or near the center of the examinee latent distribution with fewer items measuring best in the *tails* of the distribution. To the certification program's advantage, the center of the latent distribution tends to be in proximity to the passing score. In fact it is usually slightly greater than this threshold value, so that the percentage of examinees who typically pass such tests is between 50% and 80%. However, items that measure best in the tails of the distribution tend to be ignored during item selection at the passing score, thus, rarely being administered to any examinee. When periodic item-administration summaries are developed throughout a testing cycle, it becomes the psychometric staff's duty to report that

these items, usually of extremely high or low difficulty, have not been administered and are therefore of no value in the CCT.

When we are designing a CCT program for a certification agency, we consider many factors. Although we prefer to concentrate on the statistical and psychometric properties of a particular testing paradigm (i.e., those that will yield optimal testing decisions), we also have to consider political choices, such as those raised by certification directors, members of governing boards, exam panel members, and so forth. One of these latter concerns is the failure of the CCT algorithm to utilize very difficult or very easy items. The question that we are typically asked in these circumstance is "If you selected items at the ability levels of the examinees rather than at the passing score, wouldn't you use more of the item pool more efficiently and, thus also favorably impact item exposure rates?" In short, "Wouldn't this process improve the item exposure rates by *spreading around* the exposure of more items to more examinees?"

Therefore, the purpose of the current study was to investigate the effect of item selection on item exposure rates. We hypothesized that there would be significant differences in item exposure rates for the extreme items in the pool (i.e., those that measured best at high and low values of θ), but that the overall impact of these differences would be negligible.

There are several methods available to *score* and *terminate* a CCT. Normally , ACT uses a procedure called the *Sequential Probability Ratio Test* or SPRT procedure which has been described previously (Reckase, 1983; Spray & Reckase, 1987; Spray & Reckase, 1996). Another popular method of determining classification status is the direct estimation of an examinee's latent ability from the item responses within a sequential testing framework and

the comparison of that estimate, $\hat{\theta}$, in some fashion to the latent passing score, θ_p (e.g. Owen's sequential Bayes (Owen, 1975) procedure or more accurately, restricted Bayesian updating). Spray and Reckase (1996) compared the SPRT and Owen procedures directly and determined that, for a test that was unconstrained by either length or content categories, SPRT yielded a more powerful test of a single decision than the sequential Bayes procedure when items were selected to maximize information at θ_p. Although it is possible to compare the two procedures directly, as was done in the Spray and Reckase paper, it is not a simple task. Therefore, for the present study, the **only** method used to score and terminate the CCTs was the sequential Bayes procedure, so that only the **location** of item selection was manipulated (i.e., either at θ_p or $\hat{\theta}$). This ensured that the results would solely be influenced by the item-selection site and not by the method of scoring or termination.

Study Description

A simulation study was designed using an actual item pool of 1235 items. All items within the pool had been previously calibrated from item responses collected on paper/pencil administrations using the computer program *Bilog* and a 3-PL model. All items were scaled to a base form using the procedure described in Stocking and Lord (1983). Each paper/pencil form of the examination had been administered on two occasions to two separate groups of examinees. Analysis from the scaling procedure revealed that the two groups were distinctly different on their latent ability distributions. The more able group performed at a mean θ level of 0 (σ = 1), while the less able group had a mean θ = -.5 (σ = .9).

To simulate the latent distribution of examinees, a mixed normal density function was

used for all simulations. The mixed normal density function was $f(\theta) = \alpha g_1(\theta) + (1-\alpha)g_2(\theta)$, where the mixing proportion, α, was equal to .70, representing the proportion of examinees in the total sample who were more able. A mixed normal density resulted in an overall latent distribution that was slightly positively skewed (see Figure 1) with a mean equal to -.15 and variance, .9945.

Five content domains had been defined from the blueprint. The five domains required items to be administered according to a content distribution of .15, .40, .05, .20, and .20. The items in the pool had a mean expected *P*-value of .70, and the distribution of expected *P*-values was negatively skewed (see Figure 2). The passing score for this examination had been previously established as 67% correct which corresponded to a latent passing score, $\theta_p = -.69$. A graph of each item's maximum information as a function of θ shows the typical pattern that we described earlier (see Figure 3). The items tended to measure *best* at the center of the distribution and near the passing score, a pattern which repeated itself for each of the five content domains. However, the term *best* was a relative one. Most items had fairly low information, even those items that were ranked highest on information.

Test length was variable but minimum and maximum test lengths were arbitrarily set at 80 items and 120 items respectively [1]. Five thousand examinees were randomly selected from the mixed normal distribution, $f(\theta)$, to take the simulated CCT. The same examinees took each type of CCT, designated as either $CCT(\theta_p)$ or $CCT(\hat{\theta})$, depending on where the items were selected.

[1]The original paper/pencil version of the examination was 240 items in length.

Classification decisions were made for an examinee either by normal test termination or by forced classification. Normal termination occurred whenever the (1-.05) *credibility interval*, centered at $\hat{\theta}$, was either greater than or less than θ_p. A forced classification had to be made if an examinee's test had not terminated after being administered the 120th item. Forced classification was made by evaluating the most recent update of the examinee's estimated ability against the value of θ_p = -.69.

Conditional item exposure parameters were established using the Sympson-Hetter procedure (Sympson & Hetter, 1985). This procedure adjusts the exposure rate based on the rate at which items are selected by employing the item selection criteria employed in the computer simulation. The exposure control set for each item is designed to administer items such that the observed exposure rate is close to the target exposure rate. The target exposure rate (TER) of any item was set at either .20 or .10. Content constraints were controlled using a modified penalty function technique described by Swanson and Stocking (1993).

Results

Our normal interest for CCT simulations is in outcome variables such as passing rates, false positive and false negative classification rates, average test length, and so on. However, in this study, our main focus was on item exposure rates, rather than the usual variables of interest. The observed item exposure rates under the two target exposure rate (TER) conditions are described below.

TER = .20

When TER was set at .20, only 583 of the 1235 items (or 47.2%) were administered to any of the 5,000 simulated examinees under CCT($\hat{\theta}$), while only 505 items (40.9%) were used for CCT(θ_p). The average exposure rate for all items in the pool under CCT($\hat{\theta}$) was .0749 and under CCT(θ_p) was .0753.

To observe which items differed, in terms of exposure under the two item selection methods, we plotted the *difference* between individual item exposure rates and θ_{max}. Figure 4 illustrates these differences clearly, and they were not all that surprising. Under CCT($\hat{\theta}$), 64 items had item exposure rates greater than .05 from those that they had under CCT(θ_p). Conversely, under CCT(θ_p), 79 items had item exposures rates greater than .05 from those under CCT($\hat{\theta}$). Although not equal, the numbers tended to cancel each other out and, in the end, the overall exposure rates were about the same under each condition.

A similar graph of item exposure rate *differences* as a function of expected *P*-value appears in Figure 5. Although conveying the same message, this graph illustrates the item exposure picture through the eyes of the certification client who tends to understand an item's unconditional or overall difficulty better than an item's maximum information. It is this graph that the client would use to argue the selection of items at individual estimates of ability, because they would assume that by using more difficult items, other items would be exposed less often. This graph would show that, although this is true, other items at the easier end of the difficulty scale will be administered more frequently under CCT(θ_p) and these will tend to cancel each other out, resulting in about the same item exposure rates for

items in the pool.

TER = .10

When TER was set at .10, 1018 of the 1235 items (or 82.4%) were administered to any of the 5,000 simulated examinees under CCT($\hat{\theta}$), while 997 items (80.7%) were used for CCT(θ_p). The average exposure rate for all items in the pool under CCT($\hat{\theta}$) was .0777 and under CCT(θ_p) was .0782. Once again, we plotted the *difference* between individual item exposure rates for CCT(θ_p) and CCT($\hat{\theta}$) as a function of θ_{max} (see Figure 6). Under CCT($\hat{\theta}$), only 28 items had item exposure rates greater than .05 from those that they had under CCT(θ_p), while only 36 had item exposures rates greater than .05 from those under CCT($\hat{\theta}$). Once again, the numbers tended to cancel each other out which resulted in similar overall exposure rates. The plot of exposure rate differences as a function of expected *P*-value showed that very difficult items were rarely administered under CCT(θ_p), but easy items were administered at about the same frequency under both methods of item selection.

Other Outcome Results

Both methods of item selection produced about the same degree of classification accuracy. Table 1 shows the results under the TER = .20 condition, while Table 2 gives comparable results for TER = .10. Thus, both item selection algorithms appeared to produce tests of equal accuracy for about the same test length, and content constraints were met fairly well under each method. Performance of the CCTs, in terms of classification accuracy, declined slightly for TER = .10 as opposed to TER = .20, which is expected whenever we are forced to use more items with less information.

We would expect an improvement in classification accuracy with a pool of items that had a greater amount of information at the passing score than the current pool exhibited. In that case, the increased information at θ_p might yield better decisions than an equal number of high information items that measured best in the tails of the latent distribution. These examinees would tend to be classified correctly because they are so far above (or below) the passing score, and the use of highly precise items in these regions of the latent distribution would appear to be ineffective, regardless of the method of item selection.

As this study has illustrated, we would not make an argument for item selection based on improved test security. Now that we know there is little impact on item exposure rates when items are selected at θ_p as opposed to $\hat{\theta}$, we would prefer to use the SPRT procedure for this item pool. The SPRT requires items to be selected at θ_p and should produce a CCT that is at least as accurate, and possibly more so, than the sequential Bayes methods.

References

Owen, R.J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70*, 351-356.

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), *New horizons in testing; Latent trait test theory and computerized adaptive testing* (pp. 237-255). New York: Academic Press.

Reckase, M. D., & Spray, J. A. (April,1994). *The selection of test items for decision making with a computer adaptive test.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Spray J. A., & Reckase, M. D. (1996). Comparison of SPRT and sequential Bayes procedures for classifying examinees into two categories using a computerized test. *Journal of Educational and Behavioral Statistics, 21*, 405-414.

Spray, J. A., & Reckase, M. D. (1987). *The effect of item parameter estimation error on the decisions made using the sequential probability ratio test.* (Research Report No. ONR 87-1). Iowa City, IA: American College Testing.

Sympson, J. B., & Hetter, R. D. (1985). Controlling item-exposure rates in computerized adaptive testing. *Proceedings of the 27th annual meeting of the Military Testing Association.* (pp 973-977). San Diego, CA: Navy Personnel Research and Development Center.

Swanson, L., & Stocking, M. L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement, 17*, pp. 151-166.

Table 1

Simulation Results

	TER = .20		TER = .10	
	$\hat{\theta}$	θ_p	$\hat{\theta}$	θ_p
Passing Rate(%)	71.4	71.4	72.6	72.7
Failing Rate(%)	28.6	28.6	27.4	27.3
False Positive Error Rate(%)	3.8	3.7	5.3	5.2
False Negative Error Rate(%)	3.0	2.8	3.3	3.1
Total Error Rate(%)	6.8	6.5	8.6	8.3
Average Test Length	92.6	93.0	96.5	95.9
SD Test Length	18.0	18.2	19.3	19.2

Figure 1 - Mixed Normal Distribution

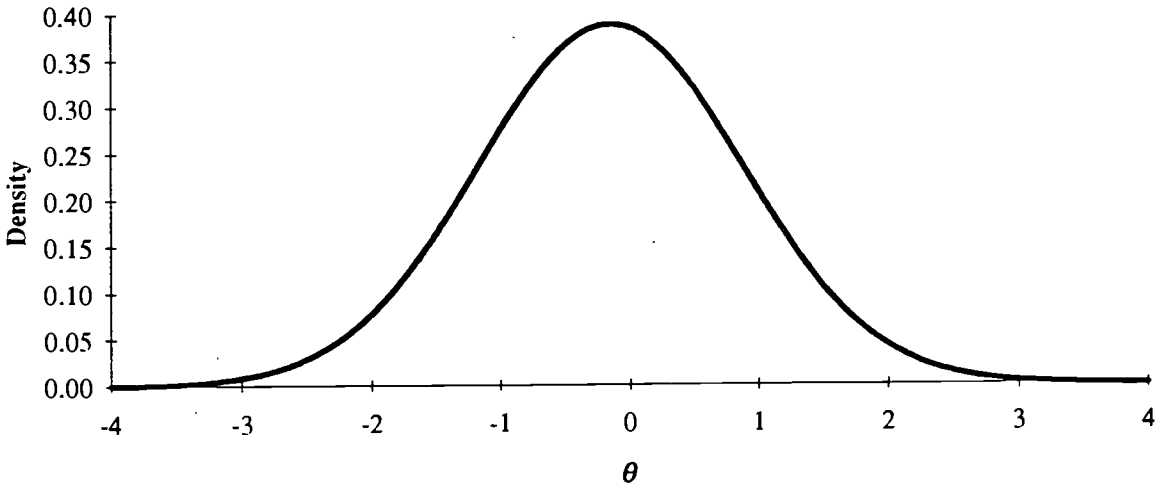


Figure 2 - Expected P-Values

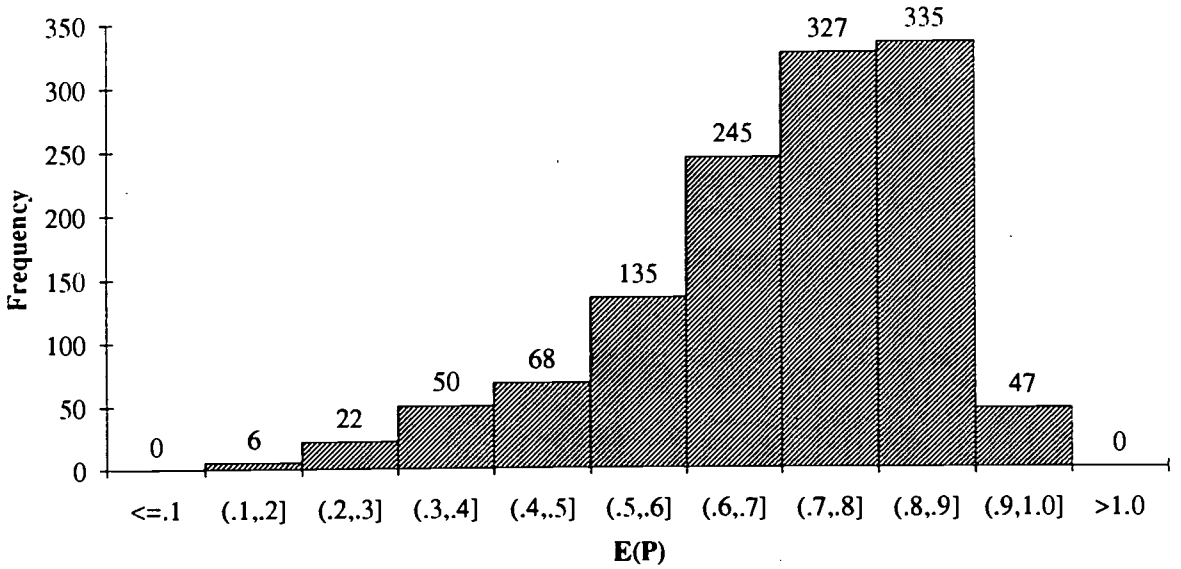


Figure 3 - Distribution of Maximum Information

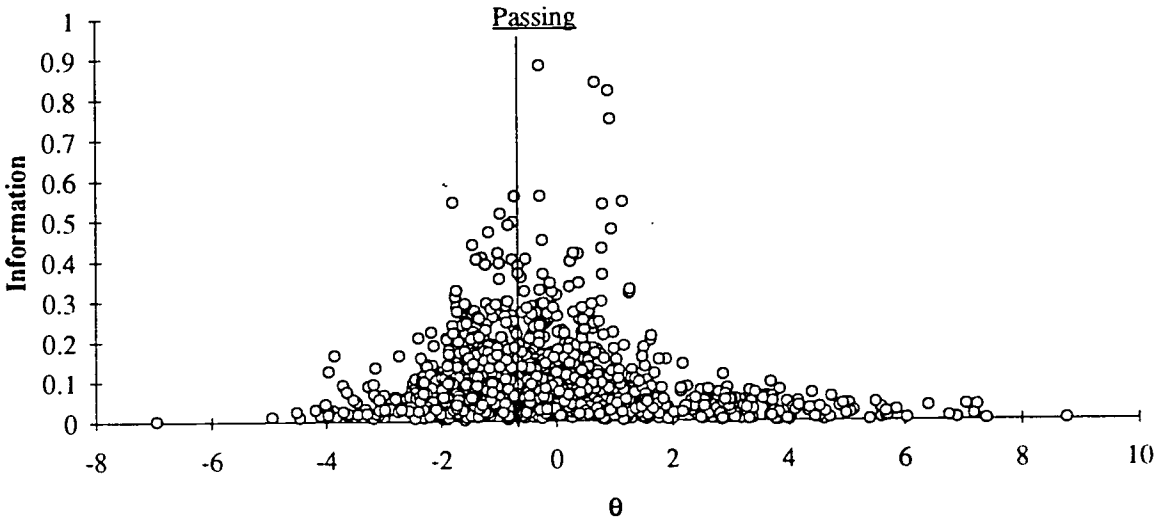


Figure 4 - Item Exposure Rate Differences ($\hat{\theta} - \theta_p$) by Maximum Item Information (TER = .20)

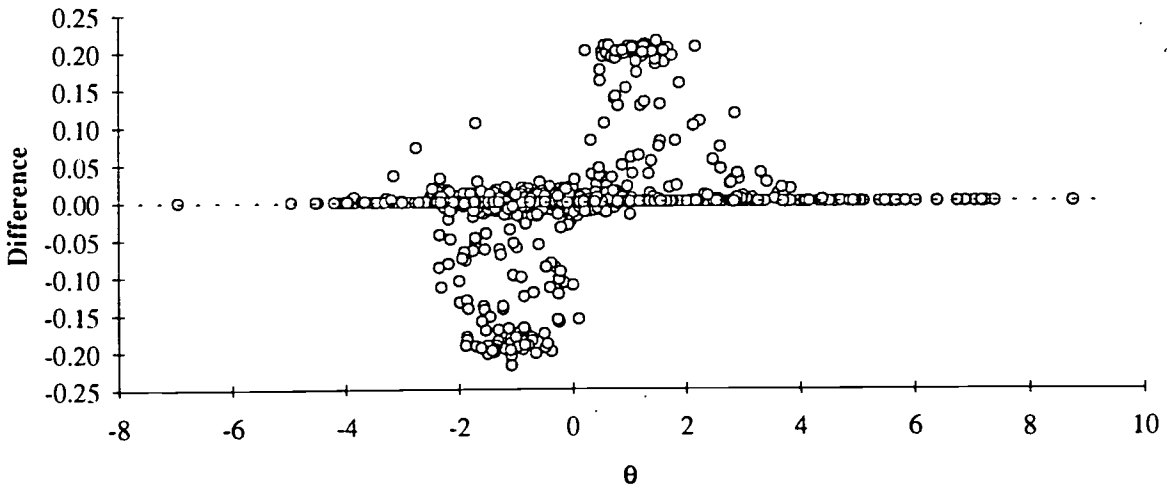


Figure 5 - Item Exposure Rate Differences ($\hat{\theta} - \theta_p$)
by Expected P-Value (TER = .20)

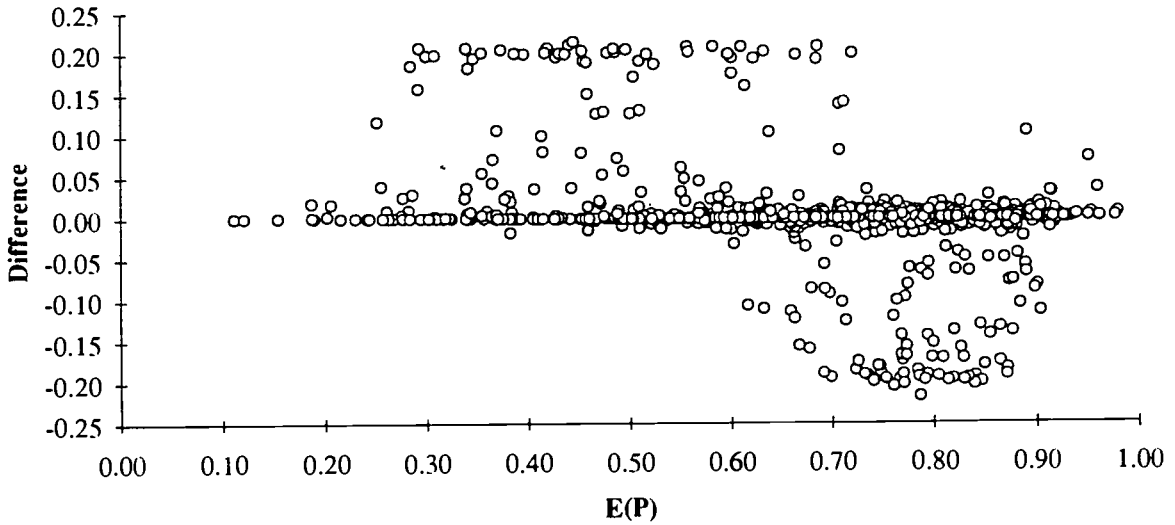


Figure 6 - Item Exposure Rate Differences ($\hat{\theta} - \theta_p$)
by Maximum Item Information (TER = .10)

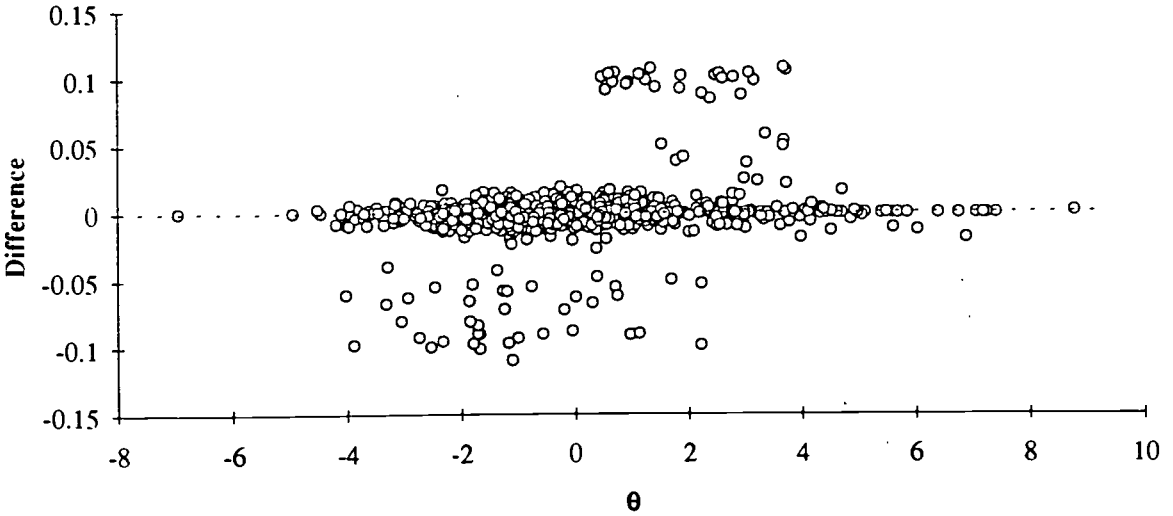
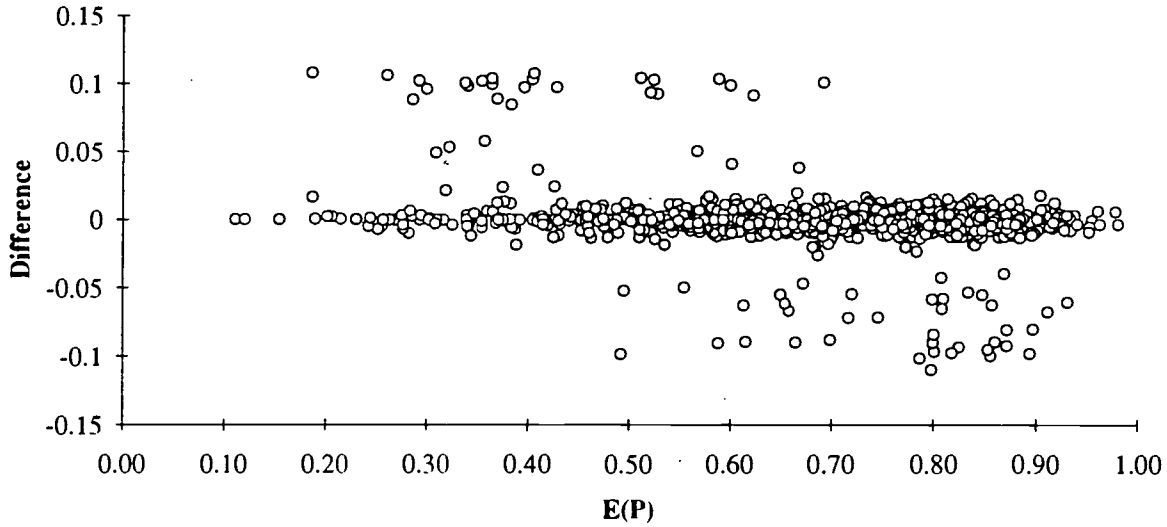


Figure 7 - Item Exposure Rate Differences ($\hat{\theta} - \theta_p$) by Expected P-Value (TER =.10)



U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM028351

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Effect of Item Selection on Item Exposure Rates Within a Computerized Classification Test.	
Author(s): John C. Kalohn & Judith A. Spray	
Corporate Source: ACT, Inc.	Publication Date: April 14, 1998

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ☑	Level 2A ☐	Level 2B ☐
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: John Kalohn	Printed Name/Position/Title: John C. Kalohn, Senior Research Assoc.
Organization/Address: ACT 2201 N. Dodge St Iowa City, IA 52243	Telephone: 319-337-1106 FAX: 319-337-1122
	E-Mail Address: Kalohn@act.org Date: 4/10/98

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

THE UNIVERSITY OF MARYLAND
ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
1129 SHRIVER LAB, CAMPUS DRIVE
COLLEGE PARK, MD 20742-5701
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com